Penalized Clustering of Large Scale Functional Data with Multiple Covariates

Ping Ma and Wenxuan Zhong *

Abstract

In this article, we propose a penalized clustering method for large scale data with multiple covariates through a functional data approach. In the proposed method, responses and covariates are linked together through nonparametric multivariate functions (fixed effects), which have great flexibility in modeling a variety of function features, such as jump points, branching, and periodicity. Functional ANOVA is employed to further decompose multivariate functions in a reproducing kernel Hilbert space and provide associated notions of main effect and interaction. Parsimonious random effects are used to capture various correlation structures. The mixed-effect models are nested under a general mixture model, in which the heterogeneity of functional data is characterized. We propose a penalized Henderson's likelihood approach for model-fitting and design a rejection-controlled EM algorithm for the estimation. Our method selects

*Ping Ma and Wenxuan Zhong are Assistant Professors (E-mail: pingma, wenxuan@uiuc.edu), Department of Statistics, University of Illinois, Champaign, IL 61820. PM's research was partially supported by National Science Foundation grant DMS-0723759. The authors are grateful to Jun S. Liu, Chong Gu, Yu Zhu, Xuming He, Steve Portnoy for many illuminating discussions on this article. The authors thank Kurt Kwast for providing the yeast microarray data. The authors also thank the editor, the associate editor, the two referees, John Marden, and Adam Martinsek for their constructive comments and suggestions that have led to significant improvement of this article.

smoothing parameters through generalized cross-validation. Furthermore, the Bayesian confidence intervals are used to measure the clustering uncertainty. Simulation studies and real-data examples are presented to investigate the empirical performance of the proposed method. Open-source code is available in the R package MFDA.

Key words: Clustering, Functional Data Analysis, Mixed-Effect Model, Smoothing Spline, EM Algorithm.

Running title: Penalized clustering of functional data.

1 Introduction

With the rapid advancement in high throughput technology, extensive repeated measurements have been taken to monitor the system-wide dynamics in many scientific investigations. A typical example is temporal gene expression studies, in which a series of micorarray experiments are conducted sequentially during a biological process, e.g., cell cycle microarray experiments (Spellman et al. 1998). At each time point, mRNA expression levels of thousands of genes are measured simultaneously. Collected over time, a gene's "temporal expression profile" gives the scientist some clues on what role this gene might play during the process. A group of genes with similar profiles are often "co-regulated" or participants of a common and important biological function. Thus clustering genes into homogeneous groups is a crucial first step to decipher the underlying mechanism. The need to account for intrinsic temporal dependency of repeated observations within the same individual renders traditional methods such as K-means and hierarchical clustering inadequate. By casting repeated observations as multivariate data with certain correlation structure, one ignores the time interval and time order of sampling. Additionally, missing observations in the measurements yield an unbalanced design, which requires imputation beforehand for application of mul-

tivariate approaches, e.g., the multivariate Gaussian clustering method (MCLUST, Fraley and Raftery 1990).

In addition to the time factor, such repeated measurements often contain other covariates, e.g., replicates at each time point, species in comparative genomics studies (McCarroll et al. 2004), and treatment groups in case-control studies (Storey et al. 2005), as well as many factors in a factorial designed experiment. Incorporation of multiple covariates adds another layer of complexity. Clustering methods taking all these factors into account are still lacking.

Recently, nonparametric analysis of data in the form of curves, i.e. functional data, is subject to active research. See Ramsay and Silverman (2005, 2002) for a comprehensive treatment of functional data analysis. A curve-based clustering method (FCM) was introduced in James and Sugar (2003) to cluster sparsely sampled functional data. Similar approaches were developed in Luan and Li (2003, 2004) and Heard et al. (2006) to analyze temporal gene expression data. Although these methods model the time factor explicitly, none of them are designed to accommodate additional factors. Moreover, smoothing-related parameters, e.g., knots and degrees of freedom, in these methods are the same across all clusters and must be specified *a priori*. Consequently, they can not model drastically different patterns among different clusters, which leads to high false classification rate. Finally, the computational costs of these methods are very high for large scale data.

Motivated by analysis of temporal gene expression data, we propose a flexible functional data clustering method that overcomes the aforementioned obstacles. In our proposed method, responses and covariates are linked together through nonparametric multivariate functions (fixed effects), which have great flexibility in modeling a variety of function features, such as jump points, branching, and periodicity. Functional ANOVA is employed to further decompose multivariate functions (fixed effects) in a reproducing kernel Hilbert space and provide associated notions of main effect

and interaction (Wahba 1990 and Gu 2002). Parsimonious random effects, complementing the fixed effects, are used to capture various correlation structures. The mixed-effect models are nested under a general mixture model, in which the heterogeneity of the functional data is characterized. We propose a penalized Henderson's likelihood approach for model-fitting and design a rejection-controlled EM algorithm for estimation. In this EM algorithm, the E-step is followed by a rejection-controlled sampling step (Liu et al. 1998) to eliminate a significant number of functional observations, whose posterior probabilities of belonging to a particular cluster is negligible, from calculation in the subsequent M-step. The M-step is decomposed into the simultaneous maximization of penalized weighted least squares in each cluster. The smoothing parameters associated with the penalty are selected by generalized cross-validation, which can be shown to track a squared error loss asymptotically. Our method is thus data-adaptive and automatically captures some important functional fluctuations. For model selection, we employ BIC to select the number of clusters. Moreover, the proposed method not only provides subject-to-cluster assignment but also the estimated mean function and associated Bayesian confidence intervals for each cluster. The Bayesian confidence intervals are used to measure the clustering uncertainty. These nice features make the proposed method extremely powerful for clustering large scale functional data.

The remainder of the article is organized as follows. In Section 2, we present a nonparametric mixed-effect model representation for functional data. A mixture model for clustering is considered in Section 3. Simulation and real data analysis follow in Section 4 and 5. A few remarks in Section 6 conclude the article. Proofs of the theorems are collected in Appendix.

2 Nonparametric Mixed-Effect Representation of Homogeneous Functional Data

Assuming the data are homogeneous, i.e., the number of clusters is one, we shall present a mixed-effect representation of functional observations.

2.1 The Model Specification

We assume the functional data of the ith individual, $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{in_i})^T$, follows the mixed-effect model,

$$\boldsymbol{y}_i = \mu(\mathbf{x}_i) + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{2.1}$$

where the population mean μ is assumed to be a smooth function defined on a generic domain Γ, $\mathbf{x}_i = (x_{i1}, \ldots, x_{in_i})^T$ is an ordered set of sampling points, $\mathbf{b}_i \sim N(0, \mathbf{B})$ is a $p \times 1$ random effect vector associated with a $n_i \times p$ design matrix $\mathbf{Z}_i$, and random errors $\boldsymbol{\epsilon}_i \sim N(0, \sigma^2\mathbf{I})$ are independent of $\mathbf{b}_i$'s, and of each other. The random effect covariance matrix $\mathbf{B}$ and random error variance σ^2 are to be estimated from the data. Model (2.1) has been extensively studied in the statistical literature. See Wang (1998), Zhang et al. (1998), Gu and Ma (2005), and references therein.

For multivariate x where $x = (x_{\langle 1\rangle}, x_{\langle 2\rangle}, \cdots, x_{\langle d\rangle})^T$, each entry $x_{\langle k\rangle}$ takes values in some fairly general domain Γ_k, i.e., $\Gamma = \otimes_{k=1}^d \Gamma_k$. Some examples are

Example 2.1 $\Gamma = [0, \mathcal{T}] \times \{1, \cdots, c\}$ to model temporal variation from time 0 to time $\mathcal{T}$ under multiple conditions; $\Gamma = \text{Circle} \times \{1, \cdots, s\}$ to model periodicity of a biological process of multiple species.

The functional ANOVA decomposition of a multivariate function μ is

$$\mu(x) = \mu_0 + \sum_{j=1}^{d} \mu_j(x_{\langle j\rangle}) + \sum_{j=1}^{d}\sum_{k=j+1}^{d} \mu_{jk}(x_{\langle j\rangle}, x_{\langle k\rangle}) + \cdots + \mu_{1,\cdots,d}(x_{\langle 1\rangle}, \cdots, x_{\langle d\rangle}) \tag{2.2}$$

where μ_0 is a constant, μ_j's are the main effects, μ_{jk}'s are the two-way interactions, and so on. The identifiability of the terms in (2.2) is assured by side conditions through averaging operators. See Wahba (1990) and Gu (2002).

By using different specifications of the random effect $\mathbf{b}_i$ and associated design matrix $\mathbf{Z}_i$, model (2.1) can accommodate various correlation structures.

Example 2.2 If we let $p = 1$, i.e., $\mathbf{b}_i$ is a scalar, $\mathbf{B} = \sigma_b^2$ and $\mathbf{Z}_i = \mathbf{1}$, we have the same correlation across time. If we let $p = 2$, i.e. $\mathbf{b}_i = (b_{i1}, b_{i2})^T$, $\mathbf{B} = \begin{pmatrix} \sigma_{b_1}^2 & \sigma_{b_1 b_2}^2 \\ \sigma_{b_1 b_2}^2 & \sigma_{b_2}^2 \end{pmatrix}$ and $\mathbf{Z}_i = (\mathbf{1}, \mathbf{x}_i)$, the difference between the ith subject profile and the mean profile is a linear function in time. The covariance between expression values at x_1 and x_2 for the same individual is $\sigma_{b_1}^2 + (x_1 + x_2)\sigma_{b_1 b_2}^2 + x_1 x_2 \sigma_{b_2}^2$.

2.2 Estimation

Model (2.1) is estimated using penalized least squares through the minimization of

$$\sum_{i=1}^{n} (\boldsymbol{y}_i - \mu(\mathbf{x}_i) - \mathbf{Z}_i \mathbf{b}_i)^T (\boldsymbol{y}_i - \mu(\mathbf{x}_i) - \mathbf{Z}_i \mathbf{b}_i) + \sum_{i=1}^{n} \sigma^2 \mathbf{b}_i^T \mathbf{B}^{-1} \mathbf{b}_i + N\lambda M(\mu), \quad (2.3)$$

for $N = \sum_i n_i$, where the first term measures the fidelity of the model to the data, $M(\mu) = M(\mu, \mu)$ is a quadratic functional that quantifies the roughness of μ, and λ is the smoothing parameter that controls the trade-off between the goodness-of-fit and the smoothness of μ. (2.3) is also referred to as penalized Henderson's likelihood since the first two terms are proportional to the joint density (Henderson's likelihood) of $(\boldsymbol{y}_i, \mathbf{b}_i)$ (Robinson 1991).

To minimize (2.3), we only need to consider smooth functions in the space $\{\mu : M(\mu) < \infty\}$ or subspace therein. As a abstract generalization of the vector spaces used extensively in multivariate analysis, Hilbert spaces inherit many nice properties

of the vector spaces. However, the Hilbert space is too loose to use for functional data analysis since even the evaluation functional $[x](f) = f(x)$, the simplest functional one may encounter, is not guaranteed to be continuous in a general Hilbert space. An example is that in the Hilbert space of square integrable functions defined on [0,1], evaluation is not even well defined. Consequently, one may focus on a constrained Hilbert space for which the evaluation functional is continuous. Such a Hilbert space is referred to as a reproducing kernel Hilbert space (RKHS), for which Ramsay and Silverman (2005) suggested a nickname: continuous Hilbert space. For example, the space of functions with square integrable second derivatives is an RKHS if it is equipped with appropriate inner products (Gu 2002). For the evaluation functional $[x](\cdot)$, by the Riesz representation theorem, there exists a nonnegative definite bivariate function $R(x, y)$, the reproducing kernel, which satisfies $\langle R(x, \cdot), f(\cdot)\rangle = f(x)$, called the "representer" of $[x](\cdot)$, in RKHS. Given an RKHS, we may derive the reproducing kernel from the Green's function associated with the quadratic functional $M(\mu)$. Since the construction of reproducing kernel is beyond the scope of this article, readers may refer to Wahba (1990) and Gu (2002) for details.

The minimization of (2.3) is performed in a reproducing kernel Hilbert space $\mathcal{H} \subseteq \{\mu : M(\mu) < \infty\}$ in which $M(\mu)$ is a square semi norm. To incorporate (2.2) in estimating multivariate functions, we consider $\mu_j \in \mathcal{H}_{\langle j\rangle}$, where $\mathcal{H}_{\langle j\rangle}$ is a reproducing kernel Hilbert space with tensor sum decomposition $\mathcal{H}_{\langle j\rangle} = \mathcal{H}_{0\langle j\rangle} \oplus \mathcal{H}_{1\langle j\rangle}$ where $\mathcal{H}_{0\langle j\rangle}$ is the finite-dimensional "parametric" subspace consisting of parametric functions, and $\mathcal{H}_{1\langle j\rangle}$ is the "nonparametric" subspace consisting of smooth functions. The induced tensor product space is

$$\mathcal{H} = \otimes_{j=1}^{d} \mathcal{H}_{\langle j\rangle} = \oplus_{\mathcal{S}}[(\otimes_{j\in\mathcal{S}} \mathcal{H}_{1\langle j\rangle}) \otimes (\otimes_{j\notin\mathcal{S}} \mathcal{H}_{0\langle j\rangle})] = \oplus_{\mathcal{S}} \mathcal{H}_{\mathcal{S}},$$

where the summation runs over all subsets $\mathcal{S} \subseteq \{1, \cdots, d\}$. These subspaces $\mathcal{H}_{\mathcal{S}}$ form

two large subspaces: $\mathcal{N}_M = \{\eta : M(\mu) = 0\}$, which is the null space of $M(\mu)$, and $\mathcal{H} \ominus \mathcal{N}_M$ with the reproducing kernel $R_M(\cdot, \cdot)$. The solution of (2.3) has an expression

$$\mu(x) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(x) + \sum_{i=1}^{T} c_i R_M(s_i, x), \tag{2.4}$$

where $\{\phi_\nu\}_{\nu=1}^m$ is a basis of $\mathcal{N}_M$, and d_ν and c_i are the coefficients, $\mathbf{s} = (s_1, \cdots, s_T)$ is a distinct combination of all $x_{ij} (i = 1, \cdots, n, j = 1, \cdots, n_i)$.

Example 2.3 Consider the temporal variation under a treatments. Take the fixed effect as $\mu(t, \tau)$, where $\tau \in \{1, \ldots, a\}$ denotes the treatment levels. One may decompose

$$\mu(t, \tau) = \mu_\emptyset + \mu_1(t) + \mu_2(\tau) + \mu_{1,2}(t, \tau),$$

where $\mu_\emptyset$ is a constant, $\mu_1(t)$ is a function of t satisfying $\mu_1(0) = 0$, $\mu_2(\tau)$ is a function of τ satisfying $\sum_{\tau=1}^a \mu_2(\tau) = 0$, and $\mu_{1,2}(t, \tau)$ satisfies $\mu_{1,2}(0, \tau) = 0$, $\forall \tau$, and $\sum_{\tau=1}^a \mu_{1,2}(t, \tau) = 0$, $\forall t$. The term $\mu_\emptyset + \mu_1(t)$ is the "average variation" and the term $\mu_2(\tau) + \mu_{1,2}(t, \tau)$ is the "contrast variation".

For flexible models, one may use

$$M(\mu) = \theta_1^{-1} \int_0^{\mathcal{T}} (d^2\mu_1/dt^2)^2 dt + \theta_{1,2}^{-1} \int_0^{\mathcal{T}} \sum_{\tau=1}^{a} (d^2\mu_{1,2}/dt^2)^2 dt, \tag{2.5}$$

which has a null space $\mathcal{N}_M$ of dimension $2a$. A set of ϕ_ν are given by

$$\{1, t, I_{\{j\}}(\tau) - 1/a, (I_{\{j\}}(\tau) - 1/a)t, j = 1, \ldots, a - 1\},$$

and the function R_M is given by

$$R_M(t_1, \tau_1; t_2, \tau_2) = \theta_1 \int_0^{\mathcal{T}} (t_1 - u)_+ (t_2 - u)_+ du + \theta_{1,2}(I_{\{\tau_1\}}(\tau_2) - 1/a) \int_0^{\mathcal{T}} (t_1 - u)_+ (t_2 - u)_+ du$$

See, e.g., Gu (2002, §2.4.4). To force an additive model

$$\mu(t,\tau) = \mu_\emptyset + \mu_1(t) + \mu_2(\tau), \tag{2.6}$$

which yields parallel curves at different treatments, one may set $\theta_{1,2} = 0$ and remove $(I_{\{j\}}(\tau) - 1/a)t$ from the list of ϕ_ν.

Substituting (2.4) into (2.3), we have

$$(\boldsymbol{y} - \mathbf{Sd} - \mathbf{Rc} - \mathbf{Zb})^T(\boldsymbol{y} - \mathbf{Sd} - \mathbf{Rc} - \mathbf{Zb}) + \mathbf{b}^T\boldsymbol{\Omega}\mathbf{b} + N\lambda\mathbf{c}^T\mathbf{Q}\mathbf{c}, \tag{2.7}$$

where $\boldsymbol{y} = (\boldsymbol{y}_1^T, \cdots, \boldsymbol{y}_n^T)^T$, $\mathbf{d} = (d_1, \cdots, d_m)^T$, $\mathbf{c} = (c_1, \cdots, c_T)^T$, $\mathbf{b} = (\mathbf{b}_1^T, \cdots, \mathbf{b}_n^T)^T$, $\mathbf{S} = (\mathbf{S}_1^T, \cdots, \mathbf{S}_n^T)^T$ with the (k,ν)th entry of the $n_i \times m$ matrix $\mathbf{S}_i$ equal to $\phi_\nu(t_{ik})$, $\mathbf{R} = (\mathbf{R}_1^T, \cdots, \mathbf{R}_n^T)^T$ with the (l,j)th entry of the $n_i \times T$ matrix $\mathbf{R}_i$ equal to $R_M(t_{il}, s_j)$, the design matrix $\mathbf{Z} = \mathrm{diag}(\mathbf{Z}_1, \cdots, \mathbf{Z}_n)$, $\boldsymbol{\Omega} = \sigma^2\mathrm{diag}(\mathbf{B}^{-1}, \cdots, \mathbf{B}^{-1})$ and $\mathbf{Q}$ is $T \times T$ matrix with the (j,k)th entry equal to $R_M(s_j, s_k)$.

Differentiating (2.7) with respect to $\mathbf{d}$, $\mathbf{c}$ and $\mathbf{b}$ and setting the derivatives to 0, one has

$$\begin{pmatrix} \mathbf{S}^T\mathbf{S} & \mathbf{S}^T\mathbf{R} & \mathbf{S}^T\mathbf{Z} \\ \mathbf{R}^T\mathbf{S} & \mathbf{R}^T\mathbf{R} + (N\lambda)\mathbf{Q} & \mathbf{R}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{S} & \mathbf{Z}^T\mathbf{R} & \mathbf{Z}^T\mathbf{Z} + \boldsymbol{\Omega} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{d}} \\ \hat{\mathbf{c}} \\ \hat{\mathbf{b}} \end{pmatrix} = \begin{pmatrix} \mathbf{S}^T\boldsymbol{y} \\ \mathbf{R}^T\boldsymbol{y} \\ \mathbf{Z}^T\boldsymbol{y} \end{pmatrix}. \tag{2.8}$$

The system (2.8) can be solved through the pivoted Cholesky decomposition followed by backward and forward substitutions. See, e.g., Kim and Gu (2004) for details.

The fitted values $\hat{\boldsymbol{y}} = \mathbf{S}\hat{\mathbf{d}} + \mathbf{R}\hat{\mathbf{c}} + \mathbf{Z}\hat{\mathbf{b}}$ of (2.3) can be written as $\hat{\boldsymbol{y}} = \mathbf{A}(\lambda, \boldsymbol{\Omega})\boldsymbol{y}$,

where $\mathbf{A}(\lambda, \mathbf{\Omega})$ is the smoothing matrix given below,

$$\mathbf{A}(\lambda, \mathbf{\Omega}) = (\mathbf{S}, \mathbf{R}, \mathbf{Z}) \begin{pmatrix} \mathbf{S}^T\mathbf{S} & \mathbf{S}^T\mathbf{R} & \mathbf{S}^T\mathbf{Z} \\ \mathbf{R}^T\mathbf{S} & \mathbf{R}^T\mathbf{R} + (N\lambda)\mathbf{Q} & \mathbf{R}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{S} & \mathbf{Z}^T\mathbf{R} & \mathbf{Z}^T\mathbf{Z} + \mathbf{\Omega} \end{pmatrix}^{+} \begin{pmatrix} \mathbf{S}^T \\ \mathbf{R}^T \\ \mathbf{Z}^T \end{pmatrix}, \tag{2.9}$$

and $\mathbf{C}^+$ denotes the Moore-Penrose inverse of $\mathbf{C}$ satisfying $\mathbf{C}\mathbf{C}^+\mathbf{C} = \mathbf{C}$, $\mathbf{C}^+\mathbf{C}\mathbf{C}^+ = \mathbf{C}^+$, $(\mathbf{C}\mathbf{C}^+)^T = \mathbf{C}\mathbf{C}^+$ and $(\mathbf{C}^+\mathbf{C})^T = \mathbf{C}^+\mathbf{C}$.

With varying smoothing parameters λ (including θ) and correlation parameters $\mathbf{\Omega}$, (2.8) defines an array of possible estimates, in which we need to choose a specific one in practice. A classic data-driven approach for selecting the smoothing parameter λ is generalized cross-validation (GCV), which was proposed in Craven and Wahba (1979). Treating the correlation parameters $\mathbf{\Omega}$ as extra smoothing parameters, we adopt the approach of Gu and Ma (2005) to estimate λ and the correlation parameters $\mathbf{\Omega}$ simultaneously through minimizing the GCV score

$$V(\lambda, \mathbf{\Omega}) = \frac{N^{-1}\boldsymbol{y}^T(\mathbf{I} - \mathbf{A}(\lambda, \mathbf{\Omega}))^2\boldsymbol{y}}{\{N^{-1}\mathrm{tr}(\mathbf{I} - \mathbf{A}(\lambda, \mathbf{\Omega}))\}^2}. \tag{2.10}$$

Since the GCV score $V(\lambda, \mathbf{\Omega})$ is non-quadratic in λ and $\mathbf{\Omega}$, one may employ standard nonlinear optimization algorithms to minimize the GCV as a function of the tuning parameters. In particular, we used the modified Newton algorithm developed by Dennis and Schnabel (1996) to find the minimizer. The distinguishing feature of generalized cross-validation is that its asymptotic optimality can be justified in a decision-theoretic framework. One may define a quadratic loss function as,

$$L(\lambda, \mathbf{\Omega}) = \frac{1}{N}\sum_{i=1}^{n}(\hat{\boldsymbol{y}}_i - \mu(\mathbf{x}_i) - \mathbf{Z}_i\mathbf{b}_i)^T(\hat{\boldsymbol{y}}_i - \mu(\mathbf{x}_i) - \mathbf{Z}_i\mathbf{b}_i).$$

Under general conditions, Gu and Ma (2005) showed that the GCV tracks the loss function asymptotically,

$$V(\lambda, \mathbf{\Omega}) - L(\lambda, \mathbf{\Omega}) - \frac{1}{N}\sum_{i=1}^{n} \boldsymbol{\epsilon}_i^T \boldsymbol{\epsilon}_i = o_p(L(\lambda, \mathbf{\Omega})).$$

Note that $\boldsymbol{\epsilon}_i$ does not depend on λ and $\mathbf{\Omega}$. It then follows that the minimizer of the GCV score $V(\lambda, \mathbf{\Omega})$ approximately minimizes the loss function $L(\lambda, \mathbf{\Omega})$.

2.3 Bayesian Confidence Intervals

Unlike confidence estimates in parametric models, a rigorously justified interval estimate is a rarity for nonparametric functional estimation. An exception is the Bayesian confidence interval developed by Wahba (1983) from a Bayes model. A nice feature of Bayesian confidence intervals is that they have a certain across-the-function coverage property. See Nychka (1988). In this section, we derive the posterior mean and variance for constructing Bayesian confidence intervals in our setting.

The regularization is equivalent to imposing a prior on the functional form of $\mu(x)$. To see this,we decompose $\mu = f_0 + f_1$, where f_0 has a diffuse prior in the space $\mathcal{N}_M$ and f_1 has an independent Gaussian process prior with mean zero and covariance,

$$E[f_1(s_k) f_1(s_l)] = \frac{\sigma^2}{N\lambda} R_M(s_k, \mathbf{s}^T) \mathbf{Q}^+ R_M(\mathbf{s}, s_l). \tag{2.11}$$

The minimizer of (2.3) can be shown to be the posterior mean under the above prior by the following theorem.

Theorem 2.1 *With the prior for μ specified above and a generic $np \times 1$ vector $\mathbf{z}$, the*

posterior mean of $\mu(x) + \mathbf{z}^T\mathbf{b}$ *has the following expression:*

$$E[\mu(x) + \mathbf{z}^T\mathbf{b}|\boldsymbol{y}] = \boldsymbol{\phi}^T\hat{\mathbf{d}} + \boldsymbol{\xi}^T\hat{\mathbf{c}} + \mathbf{z}^T\hat{\mathbf{b}}, \tag{2.12}$$

where $\boldsymbol{\phi}$ *is* $m \times 1$ *with the* ν*th entry* $\phi_\nu(x)$, $\boldsymbol{\xi}$ *is* $T \times 1$ *with the* i*th entry* $R(s_i, x)$, $\hat{\mathbf{d}}$, $\hat{\mathbf{c}}$, *and* $\hat{\mathbf{b}}$ *are the solutions of (2.8).*

The posterior variance is given in the following theorem.

Theorem 2.2 *Under the model specified in Theorem 2.1, the posterior variance has the following expression:*

$$\begin{aligned}\frac{N\lambda}{\sigma^2} Var[\mu(x) + \mathbf{z}^T\mathbf{b}|\boldsymbol{y}] &= \boldsymbol{\xi}^T\mathbf{Q}^+\boldsymbol{\xi} + N\lambda\mathbf{z}^T\boldsymbol{\Omega}^+\mathbf{z} + \boldsymbol{\phi}^T(\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S})^{-1}\boldsymbol{\phi} \\ &\quad - 2\boldsymbol{\phi}^T(\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{W}^{-1}\mathbf{R}\mathbf{Q}^+\boldsymbol{\xi} - 2N\lambda\boldsymbol{\phi}^T(\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{W}^{-1}\mathbf{Z}\boldsymbol{\Omega}^+\mathbf{z} \\ &\quad -(\boldsymbol{\xi}^T\mathbf{Q}^+\mathbf{R}^T + N\lambda\mathbf{z}^T\boldsymbol{\Omega}^+\mathbf{Z})(\mathbf{W}^{-1} - \mathbf{W}^{-1}\mathbf{S}(\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{W}^{-1})(\mathbf{R}\mathbf{Q}^+\boldsymbol{\xi} + N\lambda\mathbf{Z}\boldsymbol{\Omega}^+\mathbf{z}),\end{aligned}$$

where $\mathbf{W} = \mathbf{R}\mathbf{Q}^+\mathbf{R}^T + N\lambda\mathbf{Z}\boldsymbol{\Omega}^+\mathbf{Z}^T + N\lambda\mathbf{I}$.

The proofs of the above two theorems are given in Appendix. Using Theorem 2.1 and Theorem 2.2, we construct the $100(1-\alpha)\%$ Bayesian confidence intervals as, $E[\mu(x) + \mathbf{z}^T\mathbf{b}|\boldsymbol{y}] \pm \Phi(1-\alpha/2)^{-1}\sqrt{\mathrm{Var}[\mu(x) + \mathbf{z}^T\mathbf{b}|\boldsymbol{y}]}$, where $\Phi(1-\alpha/2)^{-1}$ is the $100(1-\alpha/2)$ percentile of the standard Gaussian distribution. Letting $\mathbf{z} = 0$, we get Bayesian confidence intervals for $\mu(x)$. Note that the construction of Bayesian confidence intervals is pointwise. It is unclear whether the across-the-function coverage property of Nychka (1988) holds in our case.

3 The Mixture Model

Based on the mixed-effect representation of homogeneous functional data, we shall now present a mixture model for characterizing the heterogeneity.

3.1 The Model Specification

When the population is heterogeneous, we assume that the ith functional observation can be modeled as

$$\boldsymbol{y}_i = \mu_k(\mathbf{x}_i) + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \qquad \text{with probability } p_k \tag{3.1}$$

where $k = 1, \cdots, K$, the kth cluster's mean μ_k is a smooth function defined on a generic domain Γ, $\mathbf{b}_i \sim N(0, \mathbf{B}_k)$ is a $p \times 1$ random effect vector associated with a $n_i \times p$ design matrix $\mathbf{Z}_i$, $\boldsymbol{\epsilon}_i \sim N(0, \sigma^2\mathbf{I})$ are random errors independent of the $\mathbf{b}_i$'s and of each other, cluster probabilities p_k satisfy $\sum_{k=1}^{K} p_k = 1$, and K is the number of clusters in the population.

To ease the computation, we introduce a "latent" membership labeling variable J_{ik} such that $J_{ik} = 1$ indicates individual i belongs to the kth cluster and $J_{ik} = 0$ otherwise. Thus we have the probability that $J_{ik} = 1$ is p_k. The mixture Henderson's likelihood is seen to be

$$\sum_{i=1}^{n} \log \sum_{k=1}^{K} [p_k f_y(\boldsymbol{y}_i; \mathbf{b}_i, J_{ik} = 1) f_b(\mathbf{b}_i; J_{ik} = 1)]$$

where f_y and f_b are probability density functions for $\boldsymbol{y}_i$ and $\mathbf{b}_i$ respectively.

3.2 Estimation

The negative penalized Henderson's likelihood of complete data $(\boldsymbol{y}_i, J_{ik})$ where $i = 1, \cdots, n$, is seen to be

$$L_c = \text{Constant} - \sum_{i=1}^{n}\sum_{k=1}^{K} J_{ik} \log p_k$$
$$+\frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{k=1}^{K} J_{ik}[(\boldsymbol{y}_i - \mu_k(\mathbf{x}_i) - \mathbf{Z}_i\mathbf{b}_i)^T(\boldsymbol{y}_i - \mu_k(\mathbf{x}_i) - \mathbf{Z}_i\mathbf{b}_i) + \sigma^2\mathbf{b}_i^T\mathbf{B}_k^{-1}\mathbf{b}_i] + \sum_{k=1}^{K} N\lambda_k M(\mu_k) \tag{3.2}$$

where λ_k is the smoothing parameter for μ_k.

Once the penalized Henderson's likelihood (3.2) is obtained, the EM algorithm (Dempster et al. 1977, Green 1990) can be derived as follows.

The E-step simply requires the calculation of

$$w_{ik} = \frac{p_k \varphi(\boldsymbol{y}_i; \mu_k(\mathbf{x}_i), \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{K} p_l \varphi(\boldsymbol{y}_i; \mu_l(\mathbf{x}_i), \boldsymbol{\Sigma}_l)} \tag{3.3}$$

where $\boldsymbol{\Sigma}_k = \mathbf{Z}_i\mathbf{B}_k\mathbf{Z}_i^T + \sigma^2\mathbf{I}$, and φ is the Gaussian density function.

The M-step requires the conditional minimization of the following equation

$$-\sum_{k=1}^{K}\sum_{i=1}^{n} w_{ik} \log p_k \tag{3.4}$$
$$+\frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}[(\boldsymbol{y}_i - \mu_k(\mathbf{x}_i) - \mathbf{Z}_i\mathbf{b}_{ik})^T(\boldsymbol{y}_i - \mu_k(\mathbf{x}_i) - \mathbf{Z}_i\mathbf{b}_{ik}) + \sigma^2\mathbf{b}_{ik}^T\mathbf{B}_k^{-1}\mathbf{b}_{ik}] + \sum_{k=1}^{K} N\lambda_k M(\mu_k), \tag{3.5}$$

where $\mathbf{b}_{ik}$ is $\mathbf{b}_i$ given the membership J_{ik}. Thus the M-step is equivalent to minimizing (3.4) and (3.5) separately.

By minimizing (3.4), we have

$$p_k = \frac{1}{n}\sum_{i=1}^{n} w_{ik} \quad \text{for} \quad k = 1, \cdots, K. \tag{3.6}$$

By minimizing (3.5), we can minimize the following K equations simultaneously

$$\sum_{i=1}^{n} w_{ik}[(\boldsymbol{y}_i - \mu_k(\mathbf{x}_i) - \mathbf{Z}_i\mathbf{b}_{ik})^T(\boldsymbol{y}_i - \mu_k(\mathbf{x}_i) - \mathbf{Z}_i\mathbf{b}_{ik}) + \sigma^2\mathbf{b}_{ik}^T\mathbf{B}_k^{-1}\mathbf{b}_{ik}] + N\lambda_k M(\mu_k) \quad k = 1, \cdots, K. \tag{3.7}$$

where $1/2\sigma^2$ is absorbed into λ_k. The minimization of (3.7) is performed in the reproducing kernel Hilbert space $\mathcal{H} \subseteq \{\eta : M(\mu) < \infty\}$. Substituting solution (2.4) into (3.7), we have

$$(\boldsymbol{y} - \mathbf{S}\mathbf{d}_k - \mathbf{R}\mathbf{c}_k - \mathbf{Z}\mathbf{b}_k)^T\mathbf{W}_k(\boldsymbol{y} - \mathbf{S}\mathbf{d}_k - \mathbf{R}\mathbf{c}_k - \mathbf{Z}\mathbf{b}_k) + \mathbf{b}_k^T\tilde{\mathbf{W}}_k^{1/2}\boldsymbol{\Omega}_k\tilde{\mathbf{W}}_k^{1/2}\mathbf{b}_k + N\lambda_k\mathbf{c}_k^T\mathbf{Q}\mathbf{c}_k, \tag{3.8}$$

where $\boldsymbol{y} = (\boldsymbol{y}_1^T, \cdots, \boldsymbol{y}_n^T)^T$, $\mathbf{d}_k = (d_{1k}, \cdots, d_{mk})^T$, $\mathbf{c}_k = (c_{1k}, \cdots, c_{Tk})^T$, $\mathbf{b}_k = (\mathbf{b}_{1k}^T, \cdots, \mathbf{b}_{nk}^T)^T$, $\mathbf{S} = (\mathbf{S}_1^T, \cdots, \mathbf{S}_n^T)^T$ with the (k, ν)th entry of the $n_i \times m$ matrix $\mathbf{S}_i$ equal to $\phi_\nu(t_{ik})$, $\mathbf{R} = (\mathbf{R}_1^T, \cdots, \mathbf{R}_n^T)^T$ with the (l, j)th entry of the $n_i \times T$ matrix $\mathbf{R}_i$ equal to $R_M(t_{il}, s_j)$, the design matrix $\mathbf{Z} = \text{diag}(\mathbf{Z}_1, \cdots, \mathbf{Z}_n)$, $\mathbf{W}_k = \text{diag}(w_{1k}\mathbf{I}_{n_1}, \cdots, w_{nk}\mathbf{I}_{n_n})$, $\tilde{\mathbf{W}}_k = \text{diag}(w_{1k}\mathbf{I}_p, \cdots, w_{nk}\mathbf{I}_p)$, $\boldsymbol{\Omega}_k = \sigma^2\text{diag}(\mathbf{B}_k^{-1}, \cdots, \mathbf{B}_k^{-1})$ and $\mathbf{Q}$ is the $T \times T$ matrix with the (j, k)th entry equal to $R_M(s_j, s_k)$.

Writing (3.8) in a more compact form, we have

$$(\boldsymbol{y}_{wk} - \mathbf{S}_{wk}\mathbf{d}_k - \mathbf{R}_{wk}\mathbf{c}_k - \mathbf{Z}\mathbf{b}_{wk})^T(\boldsymbol{y}_{wk} - \mathbf{S}_{wk}\mathbf{d}_k - \mathbf{R}_{wk}\mathbf{c}_k - \mathbf{Z}\mathbf{b}_{wk}) + \mathbf{b}_{wk}^T\boldsymbol{\Omega}_k\mathbf{b}_{wk} + N\lambda_k\mathbf{c}_k^T\mathbf{Q}\mathbf{c}_k, \tag{3.9}$$

where $\boldsymbol{y}_{wk} = \mathbf{W}_k^{1/2}\boldsymbol{y}$, $\mathbf{S}_{wk} = \mathbf{W}_k^{1/2}\mathbf{S}$, $\mathbf{R}_{wk} = \mathbf{W}_k^{1/2}\mathbf{R}$, $\mathbf{Z}_{wk} = \mathbf{W}_k^{1/2}\mathbf{Z}\tilde{\mathbf{W}}_k^{-1/2}$, and $\mathbf{b}_{wk} = \tilde{\mathbf{W}}_k^{1/2}\mathbf{b}_k$. Then (3.9) can be minimized using the techniques developed in Section 2.

The variance of measurement error is estimated as

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(\boldsymbol{y}_i - \mu_k(\mathbf{x}_i) - \mathbf{Z}_i\mathbf{b}_{ik})^T(\boldsymbol{y}_i - \mu_k(\mathbf{x}_i) - \mathbf{Z}_i\mathbf{b}_{ik}). \tag{3.10}$$

The algorithm iterates through (3.3), (3.6), (3.9) and (3.10) until all the parameters converge.

The selection of the smoothing parameters $\mathbf{\Omega}_k$ and λ_k plays an important role in the proposed algorithm. When first run our algorithm, in each iteration, the optimal smoothing parameters are selected for each cluster using GCV in (3.9). Once all parameters converge, we fixed the selected smoothing parameters and run our algorithm for fixed smoothing parameters.

After we fit the mixture model to the data, we can give a probabilistic (soft) clustering of each observation $\boldsymbol{y}_i$. That is, for each $\boldsymbol{y}_i$, $w_{i1}, \cdots, w_{iK}$ give the estimated probabilities that this observation belongs to the first, second,..., and Kth components, respectively, of the mixture. However, in many practical settings, it is highly desirable to give hard clustering of these observations by assigning each observation to one component of the mixture. In the rest of the paper, we adopt the hard clustering of McLachlan and Peel (2001) by estimating the membership label,

$$\hat{J}_{ik} = \begin{cases} 1 & \text{if} \quad k = \text{argmax}_h w_{ih} \\ 0 & \text{otherwise} \end{cases}$$

where $k = 1, \cdots, K$ and $i = 1, \cdots, n$.

3.3 Efficient Computation with Rejection Control

With thousands of observations under consideration, the E-step (3.3) results in a huge number of w_{ik}'s, many of which are extremely small. With the presence of these

small w_{ik}'s, the calculation of matrices involved in the M-step (3.9) is expensive, unstable and sometimes even infeasible. To alleviate the computation and stabilize the algorithm, we propose to add a rejection control step (Liu et al. 1998) in the EM algorithm and refer to the modified algorithm as rejection controlled EM algorithm.

Firstly, we set up a threshold value c (e.g., $c = 0.05$). Given this threshold value, we introduce the following rejection controlled step:

$$w_{ik}^* = \begin{cases} w_{ik} & \text{if} \quad w_{ik} > c \\ c & \text{with probability} \quad w_{ik}/c \quad \text{if} \quad w_{ik} \leq c \\ 0 & \text{with probability} \quad 1 - w_{ik}/c \quad \text{if} \quad w_{ik} \leq c. \end{cases}$$

The resulting w_{ik}^* needs to be normalized: $w_{ik}^{**} = w_{ik}^* / \sum_k w_{ik}^*$. Then we replace w_{ik} by w_{ik}^{**} right after the E-step (3.3). Note that when $c = 0$, the proposed algorithm is exactly the original EM algorithm, whereas the proposed algorithm reduces to a variant of Monte Carlo EM algorithm (Wei and Tanner 1990) when $c = 1$. In this way, it is possible to make accurate approximations during the E-step while greatly reducing the computation of the M-step.

Finally, in order to avoid local optima, the rejection controlled EM is run with multiple chains. In practice, we first set the threshold c close to 1 at an early stage of the iterations to expedite the calculation, then we gradually lower c so that the algorithm can achieve a better approximation of the original EM.

A critical issue arising from the new algorithm is how to choose an appropriate stopping rule. For the original EM algorithm, the likelihood function increases after each iteration, so we can stop the iteration when the likelihood does not change. However, for the rejection controlled EM algorithm, the likelihood functions fluctuates because of the sampling scheme. So a stopping rule like those used in the Gibbs sampler is employed. When the likelihood function is no longer increasing for several

consecutive iterations, we stop and choose the estimates with the highest likelihood.

3.4 The Selection of the Number of Clusters

The success of our proposed methods heavily depends on the selection of the number of clusters K. A natural choice in model-based clustering is to use the Bayesian Information Criterion (BIC). The BIC imposes a penalty on the total number of parameters, scaled by the logarithm of sample size, so as to strike a balance between the goodness-of-fit and the model complexity. A critical issue in using BIC in non-parametric settings is to determine the effective number of parameters. Here we use the trace of the smoothing matrix to approximate the number of parameters in each cluster (Hastie and Tibshirani 1990, Gu 2002). Thus BIC under our model is

$$BIC = -2\sum_{i=1}^{n}\log\sum_{k=1}^{K}p_k\varphi(\boldsymbol{y}_i;\mu_k(\mathbf{x}_i),\boldsymbol{\Sigma}_k) + (\sum_{k=1}^{K}\mathrm{tr}\mathbf{A}_k(\lambda_k,\boldsymbol{\Omega}_k) + P)\log N, \quad (3.11)$$

where $\mathbf{A}_k$ is the smoothing matrix for the kth cluster as defined in (2.9), P is the number of free parameters in p_k, λ_k, and $\boldsymbol{\Omega}_k$ where $k = 1, \cdots, K$.

4 Simulation

To assess the performance of the proposed method, we carried out extensive analysis on simulated datasets.

This simulation is designed to demonstrate the performance of the proposed method when the underlying clusters' mean functions are different for different clus-

ters. First, one hundred replicates of samples were generated according to

$$
\begin{aligned}
y_{1ij\tau} &= 3\sin(6\pi t_j)(1-t_j) + 2I_{\{1\}}(\tau) - 1 + \epsilon_{1ij\tau}, \quad i = 1, \cdots, 30; \\
y_{2ij\tau} &= 3\sin(6\pi t_j)(1-t_j) + \epsilon_{2ij\tau}, \quad i = 1, \cdots, 40; \\
y_{3ij\tau} &= 1980 t_j^7 (1-t_j)^3 + 858 t_j^2 (1-t_j)^{10} - 2 + \epsilon_{3ij\tau}, \quad i = 1, \cdots, 50; \\
y_{4ij\tau} &= 3\sin(2\pi t_j) + 2I_{\{1\}}(\tau) - 1 + \epsilon_{4ij\tau}, \quad i = 1, \cdots, 30;
\end{aligned}
$$

where $t_j = 1/15, 2/15, \cdots, 1$, $\tau = 0, 1$, indicator function $I_{\{1\}}(\tau) = 1$ if $\tau = 1$ and 0 otherwise, random errors ϵ were generated from a Gaussian distribution with mean zero and covariance matrix as follows:

$$
\begin{aligned}
&\mathrm{Var}[\epsilon_{lij\tau}] = 1, \qquad \mathrm{Cov}(\epsilon_{lij\tau_1}, \epsilon_{lik\tau_2}) = 0.2, \qquad \text{for} \quad l = 1, 3; \\
&\mathrm{Var}[\epsilon_{lij\tau}] = 1.2, \qquad \mathrm{Cov}(\epsilon_{lij\tau_1}, \epsilon_{lik\tau_2}) = 0.4, \qquad \text{for} \quad l = 2, 4;
\end{aligned}
$$

We analyzed the simulated data using the proposed method with the following mixture model

$$
\boldsymbol{y}_i = \mu_k(\mathbf{t}, \tau) + b_i \mathbf{1} + \boldsymbol{\epsilon}_i \qquad \text{with probability } p_k,
$$

where $k = 1, \cdots, K$, $\tau = 1, 2$ for two groups, $b_i \sim N(0, \sigma_b^2)$ is the individual specific random effect. The important feature of the simulated data is that the true mean curves in two groups, indexed by τ, are either identical or parallel. This information was built into our method through enforcing the additive model (2.6). The penalized Henderson's likelihood was employed for estimation with roughness penalty $M(\mu) = \int_0^1 (d^2\mu_1/dt^2)^2 dt$.

We compared our method with MCLUST (Fraley and Raftery 1990), FCM classification likelihood (FCMc), and FCM mixture likelihood (FCMm) (James and Sugar 2003).

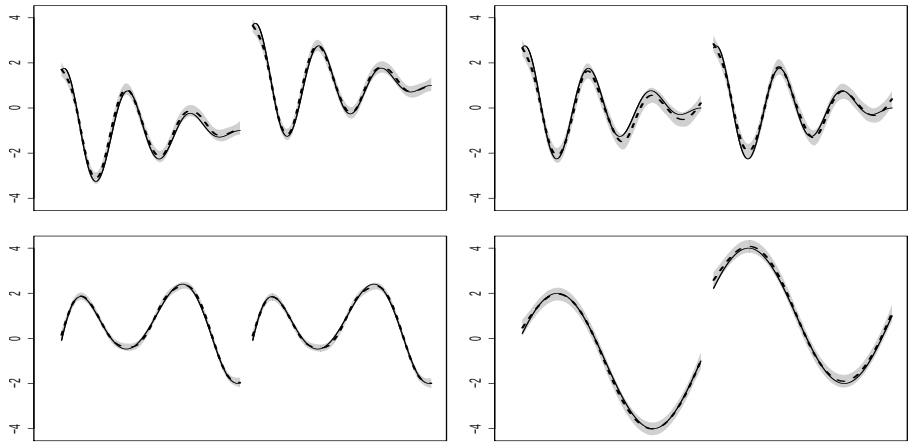

Figure 4.1: The estimated mean curves (dash lines) and 95% Bayesian confidence intervals for one simulated dataset. The true functions are superimposed as solid lines.

Since the number of clusters must be specified *a priori* in the partially implemented FCM software, we gave a significant starting advantage to the FCM algorithm by letting the number of clusters be the true number of clusters (four). For MCLUST, the clustering result with optimal BIC was reported, which was estimated from eight models with different covariance structures. The estimated mean curves using the proposed method for each cluster and the true curves of one sample are plotted in Figure 4.1.

For comparison, we need a measure of the agreement of the clustering results with the true cluster membership. A popular one is the Rand index, which is the percentage of concordance pairs over all possible data pairs. Hubert and Arabie (1985) proposed an adjusted Rand index, which takes one as the maximum value when two clustering results are the same and the expected value is equal to zero when two clustering results are independent. We found that across 100 samples the average of the adjusted Rand indices for the proposed method is 0.9676 (median is 0.9838), whereas those of MCLUST, FCMm, FCMc are 0.7553, 0.8936, and 0.8896, respectively. Moreover, the

inter-quartile range of the adjusted Rand indices of the proposed method is 0.0565, (0.0972, 0.2189 and 0.2262 for MCLUST, FCMm, and FCMc respectively). These results suggest that the proposed method outperforms FCMc and FCMm (even under the ideal scenario where the true number of clusters is provided to FCMc and FCMm a priori) as well as MCLUST.

5 Real Data Examples

5.1 Comparative Genomic Study of Fruitfly and Worm Gene Expressions

Development is an important biological process that shares many common features among different organisms. It is well-known that *D. melanogaster* (fruitfly) and *C. elegance* (worm) are two highly diverged species, the last common ancestor of which existed about one billion years ago. Their development is an active research area: In Arbeitman et al. (2002), the mRNA levels of 4028 genes in *D. melanogaster* were measured using cDNA microarrays during 62 time points starting at fertilization and spanning embryonic, larval, pupal (metamophosis) stages and the first 30 days of adulthood. mRNA was extracted from mixed male and female populations until adulthood when males and females were sampled separately. Jiang et al. (2001) reported a cDNA microarray experiment for 17871 genes over the life-cycle of *C. elegans* at 6 time points, including eggs, larval stages: L1, L2, L3 and L4, and young adults.

To study the genomic connections in expression patterns across the two species, we combined the gene expression datasets of Arbeitman et al. (2002) and Jiang et al. (2001) using the orthologous genes provided by McCarroll et al. (2004), which resulted in a merged expression dataset containing 808 orthologous genes. We analyzed the data

using the proposed method with the mixture model,

$$\boldsymbol{y}_i = \mu_k(\mathbf{t}, \tau) + b_i \mathbf{1} + \boldsymbol{\epsilon}_i$$

with probability p_k where $k = 1, \cdots, K$, $\tau = 1$ for fruitfly and $\tau = 2$ for worm, $b_i \sim N(0, \sigma_b^2)$ is the gene specific random effect. The penalized Henderson's likelihood was employed with roughness penalty M of the form (2.5). Sex differentiation of the fruitfly was modeled by a branching spline (Silverman and Wood 1987), the general analytic form of which with two branches on the right is

$$\mu(t) = \begin{cases} \sum_{\nu=1}^{m} d_\nu \phi_\nu(t) + \sum_{i=1}^{k} c_i R_M(s_i, t) & \text{if} \quad t \le s_k \\ \sum_{\nu=1}^{m} d_\nu \phi_\nu(t) + \sum_{i=1}^{k} c_i R_M(s_i, t) + \sum_{i=k+1}^{T} c_{1i} R_M(s_i - s_k, t - s_k) & \text{if} \quad t > s_k \\ \sum_{\nu=1}^{m} d_\nu \phi_\nu(t) + \sum_{i=1}^{k} c_i R_M(s_i, t) + \sum_{i=k+1}^{T} c_{2i} R_M(s_i - s_k, t - s_k) & \text{if} \quad t > s_k \end{cases}$$

where s_k is the branching point, and the second and third rows are expressions of the two branches. A cubic smoothing spline was used. The 808 genes were clustered by our method into 34 clusters. Biological functions of genes in each cluster were annotated using Gene Ontology, and Bonferroni corrected P-values of biological function enrichment were calculated based on the hypergeometric distribution (Castillo-Davis and Hartl 2003). Of the 34 clusters discovered, 21 clusters exhibit significant biological functions over-representation (P-value < 0.05). The estimated mean gene expression curves of three clusters and their 95% Bayesian confidence intervals are given in Figure 5.1.

In cluster A, which consists of 31 genes, gene expressions of worms have peaks at eggs, larva and young adult. In the same cluster, we observed that fruit-fly gene expressions that are up-regulated during embryogenesis are also up-regulated during metamorphosis, suggesting that many genes used for pattern formation during embryogenesis (the transition from egg to larva) are re-deployed during metamorphosis

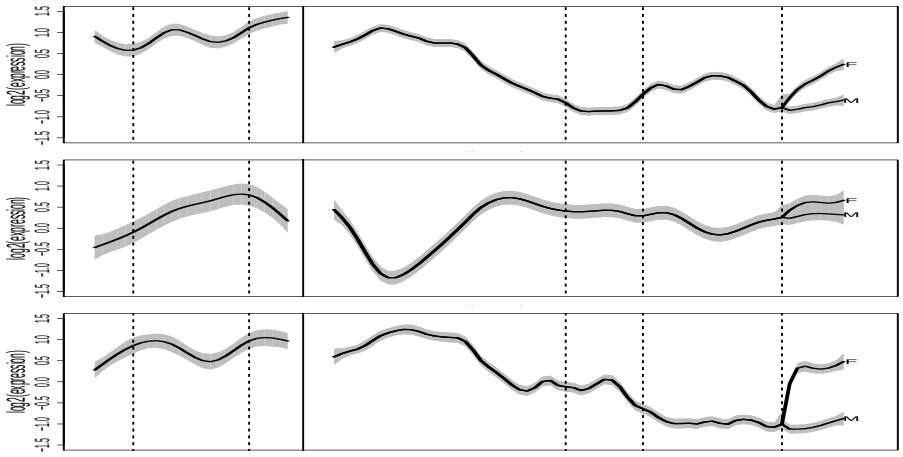

Figure 5.1: Estimated mean expression curves and 95% Bayesian confidence intervals (grey bands) for cluster A, B and C (from top to bottom) discovered in the worm-fly temporal expression data. Vertical solid lines separate worm (eggs, larva and young adult are separately by dash lines in the left frame), fruit-fly (embryogenesis, larva, pupa, and adult stages are separated by dash lines in the right frame). Adult fruit-fly male and female mean expression curves are labeled as M and F, respectively.

(the transition from larva to adult). Consistently, this cluster is enriched for genes involved in embryonic development (P-value =0.0003), post-embryonic body morphogenesis (P-value =0.007), and mRNA processing (P-value = 0.002), among others.

In cluster B, consisting of 24 genes, gene expressions of worms increase starting at eggs until they reach a peak at late larval stage. Then expressions go down during adulthood. However, we observed that fruit-fly gene expressions that are down-regulated during embryogenesis are up-regulated during metamorphosis and adult, suggesting that many genes are involved in development. The enriched gene functions are embryonic (P-value =0.02), larval development (P-value =0.008), and growth regulation (P-value $< 10^{-5}$).

Cluster C contains 25 genes. For worms, gene expressions have peaks at larva

and adult stages. An over-representation of gene functions such as reproduction (P-value $< 10^{-6}$), larval development (P-value $< 10^{-7}$). Fruit-flies show peaks in gene expression in the early embryo, and older females (but not males). An over-representation of gene functions such as reproduction (P-value $< 10^{-6}$) and embryonic development(P-value $< 10^{-5}$) are present in this cluster. Among related functions, this cluster also contains functions of female gamete generation, growth, and positive regulation of growth rate. Genes of this cluster are thus inferred to participate in sex determination, female production of eggs, and growth regulation.

5.2 Budding Yeast Gene Expression under Aerobic and Anaerobic Conditions

To study the oxygen-responsive gene networks, Lai et al. (2006) used cDNA microarray to monitor the gene expression changes of wild-type budding yeast (*Saccharomyces cerevisiae*) under aerobic and anaerobic conditions in a galactose medium. Under the aerobic conditions, the oxygen concentration was lowered gradually until oxygen was exhausted during a period of ten minutes. After 24 hours of anaerobiosis, the oxygen concentration was progressively increased back to normal level during another period of ten minutes, which was referred to as the anaerobic conditions. Microarray experiments were conducted at 14 time points under aerobic conditions and 10 time points under anaerobic conditions. A reference sample pooled from all time points was used for hybridization.

For their analysis, Lai et al. (2006) normalized gene expressions to gene expressions of time 0 and filtered out differentially expressed genes. Thus the normalized expressions at 23 time points of 2388 differentially expressed genes are used for our clustering analysis. We modeled normalized gene expression $\boldsymbol{y}_i$ of the ith gene using

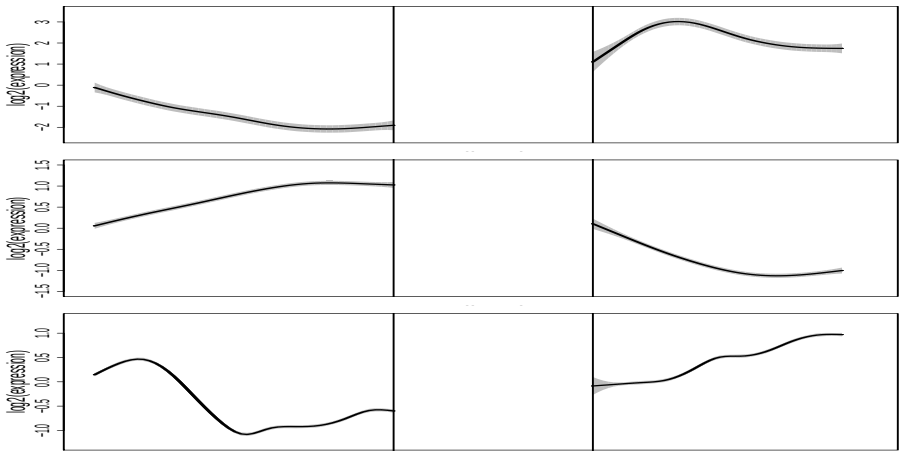

Figure 5.2: Estimated mean expression curves and 95% Bayesian confidence intervals (grey bands) for cluster A, B and C (from top to bottom) discovered in the yeast aerobic and anaerobic expression data. The aerobic (left) and anaerobic (right) conditions were separated by two vertical lines.

the mixture model,

$$\boldsymbol{y}_i = \mu_k(\mathbf{t}, \tau) + b_i \mathbf{1} + \boldsymbol{\epsilon}_i \quad \text{with probability } p_k$$

where $k = 1, \cdots, K$, $\tau = 1$ for aerobic and $\tau = 2$ for anaerobic condition, $b_i \sim N(0, \sigma_b^2)$ is the gene specific random effect. We fit the model using the penalty (2.5) with $a = 2$. In total, 2388 genes were clustered into 28 clusters using our method. FunSpec (Robinson et al. 2002) was used for gene annotation and biological function enrichment analysis. We found 26 clusters out of 28 clusters discovered have over-represented biological functions. The estimated mean gene expression profiles and associated Bayesian confidence intervals of three clusters are given in Figure 5.2.

In cluster A, which consists of 57 genes, the estimated mean expression goes down progressively as oxygen level goes down, which suggests that the genes in this cluster are transiently down-regulated in response to anaerobisis. Furthermore, the estimated

mean expression increases as oxygen concentration shifts back to normal level. Accordingly, genes involved in respiration, lipid fatty-acid and isoprenoid biosynthesis, and cell defense are over-represented in this cluster (P-value $\leq 10^{-5}$).

In contrast to cluster A, cluster B (85 genes) consists of genes involved in various biosynthesis, metabolism and catabolism such as glucose metabolism (P-value$\leq$ 10^{-6}). These biological processes are necessary to maintain the basic living needs of yeast cells. Interestingly, the alcohol biosythesis and metabolism are also enriched in this cluster. Consistent with biological function over-representation, the estimated mean expression is up-regulated in aerobic conditions and down-regulated in anaerobic conditions.

We have 70 genes in cluster C, where the estimated mean gene expression goes up at the beginning and then drops down rapidly under aerobic conditions. Under anaerobic conditions, the estimated mean gene expression is up-regulated. In this cluster, respiratory deficiency and carbon utilization are also over-represented (P-value$\leq 10^{-8}$). The initial up-regulation of gene expression under aerobic conditions can be partly explained by the fact that the cell increases energy up-taking through other biological processes, such as carbon utilization, when oxygen goes down. But as the oxygen level continues to drop, these processes are replaced by more energy efficient processes, such as glucose metabolism. Under the anaerobic conditions, these processes are revitalized again as oxygen level increases.

6 Discussion

In this article, we propose a clustering method for large scale functional data with multiple covariates. Nonparametric mixed-effect models were built, which were nested under a mixture model. The penalized Henderson's likelihood was employed for estimation. Data-driven smoothing parameters, selected through generalized cross-

validation, were used to automatically capture the functional features. The rejection-controlled EM algorithm was designed to reduce the expensive computational cost for large scale data. The simulation analyses suggest that the proposed method outperforms the existing clustering methods. Moreover, the Bayesian interpretation of the proposed method allows the development of an equivalent fully Bayesian functional data clustering method, which can accommodate additional genomic and proteomic information for gene expression study. Although it was motivated for clustering temporal expression data, our proposed method has a wide spectrum of applications, including those involving seismic wave data arising from geophysical research (Wang et al. 2006 and Ma et al. 2007). The calculations reported in this article were performed in R. Open-source code is available in the R package MFDA.

As a sequel to this work, a clustering method for discrete data, especially those arising from temporal text mining, is under active development.

Appendix

Proof of Theorem 2.1 :

Note the fact that if we specify the prior for f_0 as a Gaussian process with mean zero and covariances $E[f_0(s_k)f_0(s_l)] = \tau^2 \sum_{\nu=1}^{m} \phi_\nu(s_k)\phi_\nu(s_l)$, then when $\tau^2 \to \infty$, the prior for f_0 becomes a diffuse prior; see Wahba (1983) and Gu (2002).

Assuming $f_0(t)$ has a Gaussian process prior specified above, $f_1(x)$ has a Gaussian process prior specified as in Theorem 2.1, and $\mathbf{b}$ follows a normal distribution with mean zero and variance-covariance matrix $\mathbf{B}$, we can derive that the joint distribution of $\boldsymbol{y}$ and $f_0(x)+f_1(x)+\mathbf{z}^T\mathbf{b}$ follows a Gaussian distribution with mean 0 and covariance matrix

$$\begin{pmatrix} b\mathbf{F}\mathbf{V}^+\mathbf{F}^T + \tau^2\mathbf{S}\mathbf{S}^T + \sigma^2\mathbf{I} & b\mathbf{F}\mathbf{V}^+\tilde{\boldsymbol{\xi}} + \tau^2\mathbf{S}\boldsymbol{\phi} \\ b\tilde{\boldsymbol{\xi}}^T\mathbf{V}^+\mathbf{F}^T + \tau^2\boldsymbol{\phi}^T\mathbf{S}^T & b\tilde{\boldsymbol{\xi}}^T\mathbf{V}^+\tilde{\boldsymbol{\xi}} + \tau^2\boldsymbol{\phi}^T\boldsymbol{\phi} \end{pmatrix} \tag{6.1}$$

where $\tilde{\boldsymbol{\xi}} = (R_1(s_1, x), \ldots, R_1(s_T, x), \mathbf{z})^T$ is $(T + p) \times 1$, $\boldsymbol{\phi}$ is $m \times 1$ with the νth entry $\phi_\nu(t)$, $\mathbf{F} = (\mathbf{R}, \mathbf{Z})$, and V^+ is the Moore-Penrose inverse of $\mathbf{V} = \mathrm{diag}(\mathbf{Q}, \frac{1}{N\lambda}\boldsymbol{\Omega})$ satisfying $\mathbf{V}\mathbf{V}^+\mathbf{F}^T = \mathbf{F}^T$.

Standard calculation yields

$$\begin{aligned} E[\mu(x) + \mathbf{z}^T\mathbf{b}|\boldsymbol{y}] &= (b\tilde{\boldsymbol{\xi}}^T\mathbf{V}^+\mathbf{F}^T + \tau^2\boldsymbol{\phi}^T\mathbf{S}^T)(b\mathbf{F}\mathbf{V}^+\mathbf{F}^T + \tau^2 SS^T + \sigma^2\mathbf{I})^{-1}\boldsymbol{y} \\ &= \rho\boldsymbol{\phi}^T\mathbf{S}^T(\mathbf{W} + \rho\mathbf{S}\mathbf{S}^T)^{-1}\boldsymbol{y} + \tilde{\boldsymbol{\xi}}^T\mathbf{V}^+\mathbf{F}^T(\mathbf{W} + \rho\mathbf{S}\mathbf{S}^T)^{-1}\boldsymbol{y}, \end{aligned}$$

where $\rho = \tau^2/b$, $N\lambda = \sigma^2/b$, and $\mathbf{W} = \mathbf{F}\mathbf{V}^+\mathbf{F}^T + N\lambda\mathbf{I}$. Now letting $\rho \to \infty$, we have

$$\lim_{\rho\to\infty}(\rho\mathbf{S}\mathbf{S}^T + \mathbf{W})^{-1} = \mathbf{W}^{-1} - \mathbf{W}^{-1}\mathbf{S}(\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{W}^{-1}, \tag{6.2}$$

$$\lim_{\rho\to\infty}\rho\mathbf{S}^T(\rho\mathbf{S}\mathbf{S}^T + \mathbf{W})^{-1} = (\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{W}^{-1}. \tag{6.3}$$

See Wahba (1983) and Gu (2002) for the proof.

Therefore, $\lim_{\tau^2\to\infty} E[\mu(x) + \mathbf{z}^T\mathbf{b}|\boldsymbol{y}] = \boldsymbol{\phi}^T\mathbf{d} + \tilde{\boldsymbol{\xi}}^T\tilde{\mathbf{c}}$, where

$$\mathbf{d} = (\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{W}^{-1}\boldsymbol{y}, \tilde{\mathbf{c}} = \mathbf{V}^+\mathbf{F}^T(\mathbf{W}^{-1} - \mathbf{W}^{-1}\mathbf{S}(\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{W}^{-1})\boldsymbol{y}. \tag{6.4}$$

It is straightforward to verify that the $\mathbf{d}$ and $\tilde{\mathbf{c}}$ given in (6.4) satisfy (2.8).

Proof of Theorem 2.2 :

The posterior variance can be easily calculated by using expression (6.1) as follows,

$$\mathrm{var}[\mu(x)+\mathbf{z}^T\mathbf{b}|\boldsymbol{y}] = \tilde{\boldsymbol{\xi}}^T\mathbf{V}^+\tilde{\boldsymbol{\xi}}+\rho\boldsymbol{\phi}^T\boldsymbol{\phi}-(\tilde{\boldsymbol{\xi}}^T\mathbf{V}^+\mathbf{F}^T+\rho\boldsymbol{\phi}^T\mathbf{S}^T)(\mathbf{W}+\rho\mathbf{S}\mathbf{S}^T)^{-1}(\mathbf{F}\mathbf{V}^+\tilde{\boldsymbol{\xi}}+\rho\mathbf{S}\boldsymbol{\phi})$$

Notice that $\lim_{\rho\to\infty}\rho I-\rho^2\mathbf{S}^T(\rho\mathbf{S}\mathbf{S}^T+\mathbf{W})^{-1}\mathbf{S} = (\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S})^{-1}$, and $\mathbf{V}\mathbf{V}^+\mathbf{F}^T = \mathbf{F}^T$.

Therefore as $\rho \to \infty$, we have

$$\lim_{\tau^2 \to \infty} \mathrm{Var}[\mu(x) + \mathbf{z}^T\mathbf{b} | \boldsymbol{y}]/b = \tilde{\boldsymbol{\xi}}^T \mathbf{V}^+ \tilde{\boldsymbol{\xi}} + \boldsymbol{\phi}^T (\mathbf{S}^T \mathbf{W}^{-1} \mathbf{S})^{-1} \boldsymbol{\phi} - 2\boldsymbol{\phi}^T (\mathbf{S}^T \mathbf{W}^{-1} \mathbf{S})^{-1} \mathbf{S}^T \mathbf{W}^{-1} \mathbf{F} \mathbf{V}^+ \tilde{\boldsymbol{\xi}} - \tilde{\boldsymbol{\xi}}^T \mathbf{V}^+ \mathbf{F}^T (\mathbf{W}^{-1} - \mathbf{W}^{-1} \mathbf{S} (\mathbf{S}^T \mathbf{W}^{-1} \mathbf{S})^{-1} \mathbf{S}^T \mathbf{W}^{-1}) \mathbf{F} \mathbf{V}^+ \tilde{\boldsymbol{\xi}}.$$

References

Arbeitman, M., E. Furlong, F. Imam, E. Johnson, B. H. Null, B. S. Baker, M. Krasnow, M. P. Scott, R. W. Davis, and K. P. White (2002). Gene expression during the life cycle of drosophila melanogaster. *Science 297*(5590), 2270–2275.

Castillo-Davis, C. and D. Hartl (2003). Genemerge: post-genomic analysis, data-mining and hypothesis. *Bioinformatics 19*, 891–892.

Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math. 31*, 377–403.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B 39*, 1–37 (with discussions).

Dennis, J. E. and R. B. Schnabel (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Philadelphia: SIAM. Corrected reprint of the 1983 original.

Fraley, C. and A. E. Raftery (1990). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc. 97*, 611–631.

Green, P. J. (1990). On the use of the EM algorithm for penalized likelihood estimation. *J. Roy. Statist. Soc. Ser. B 52*, 443–452.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. New York: Springer-Verlag.

Gu, C. and P. Ma (2005). Optimal smoothing in nonparametric mixed effect models. *Ann. Statist. 33*, 1357–1379.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. London: Chapman & Hall.

Heard, N. A., C. C. Holmes, and D. A. Stephens (2006). A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of bayesian hierarchical clustering of curves. *J. Amer. Statist. Assoc. 101*(473), 18–29.

Hubert, L. and P. Arabie (1985). Comparing partitions. *J. Classification 2*, 193–218.

James, G. M. and C. A. Sugar (2003). Clustering for sparsely sampled functional data. *J. Amer. Statist. Assoc. 98*(462), 397–408.

Jiang, M., J. Ryu, M. Kiraly, K. Duke, V. Reinke, and S. K. Kim (2001). Genome-wide analysis of developmental and sex-regulated gene expression profiles in caenorhabditis elegans. *Proc. Natl. Acad. Sci. 98*(1), 218–223.

Kim, Y.-J. and C. Gu (2004). Smoothing spline gaussian regression: More scalable computation via efficient approximation. *J. Roy. Statist. Soc. Ser. B 66*, 337–356.

Lai, L. C., A. L. Kosorukoff, P. Burke, and K. E. Kwast (2006). Metabolic-state-dependent remodeling of the transcriptome in response to anoxia and subsequent reoxygenation in saccharomyces cerevisiae. *Eukaryot Cell 5*, 1468–89.

Liu, J. S., R. Chen, and W. H. Wong (1998). Rejection control and sequential importance sampling. *J. Amer. Statist. Assoc. 93*, 1022–1031.

Luan, Y. and H. Li (2003). Clustering of time-course gene expression data using a mixed-effects models with B-spline. *Bioinformatics 19*(4), 474–282.

Luan, Y. and H. Li (2004). Model-based methods for identifying periodically regulated genes based on the time course microarray gene expression data. *Bioinformatics 20*(4), 332–339.

Ma, P., P. Wang, L. Tenorio, M. V. de Hoop, and R. D. van der Hilst (2007). Imaging of structure at and near the core mantle boundary using a generalized Radon transform: 2. statistical inference of singularities. *J. Geophys. Res. 112*, B08303.

McCarroll, S. A., C. T. Murphy, S. Zou, S. D. Pletcher, C. S. Chin, Y. N. Jan, C. Kenyon, C. I. Bargmann, and H. Li (2004). Comparing genomic expression patterns across species identifies shared transcriptional program in aging. *Nature Genetics 36*(2), 197–204.

McLachlan, G. J. and D. Peel (2001). *Finite Mixture Models*. John Wiley & Sons.

Nychka, D. (1988). Bayesian confidence intervals for smoothing splines. *J. Amer. Statist. Assoc. 83*, 1134–1143.

Ramsay, J. O. and B. W. Silverman (2002). *Applied Functional Data Analysis: Methods and Case Studies*. Springer-Verlag Inc.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis*. Springer-Verlag Inc.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of the random effects. *Statist. Sci. 6*, 15–51 (with discussions).

Robinson, M. D., J. Grigull, N. Mohammad, and T. R. Hughes (2002). Funspec: a web-based cluster interpreter for yeast. *BMC Bioinformatics 3*, 3–35.

Silverman, B. W. and J. T. Wood (1987). The nonparametric estimation of branching curves. *J. Amer. Statist. Assoc. 82*, 551–558.

Spellman, P. T., G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, and B. Botstein D Futcher (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell. 9*(12), 3273–97.

Storey, J. D., W. Xiao, J. T. Leek, and R. Tompkins, R. G.and Davis (2005). Significance of time course microarray experiments. *Proc. Natl. Acad. Sci. 102*, 12837–12842.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B 45*, 133–150.

Wahba, G. (1990). *Spline Models for Observational Data*, Volume 59 of CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia: SIAM.

Wang, P., M. V. de Hoop, R. D. van der Hilst, P. Ma, and L. Tenorio (2006). Imaging of structure at and near the core mantle boundary using a generalized radon transform: 1. construction of image gathers. *J. Geophys. Res. 111*, B12304.

Wang, Y. (1998). Mixed-effects smoothing spline ANOVA. *J. Roy. Statist. Soc. Ser. B 60*, 159–174.

Wei, G. C. and M. A. Tanner (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc. 85*, 699–704.

Zhang, D., X. Lin, J. Raz, and M. Sowers (1998). Semiparametric stochastic mixed models for longitudinal data. *J. Amer. Statist. Assoc. 93*, 710–719.